An Expert-Guided Decision Tree Construction Strategy: An Application in Knowledge Discovery with Medical Databases

Yuh-Show Tsai, M.S.,[+] Paul H. King, Ph.D.,[+*]
Michael S. Higgins, M.D..[*] Donald Pierce, M.D., Ph.D.,[*] Nimesh P. Patel, M.S.[*]
[+]Department of Biomedical Engineering, [*]Department of Anesthesiology
Vanderbilt University, Nashville, Tennessee

Abstract: *With the steady growth in electronic patient records and clinical medical informatics systems, the data collected for routine clinical use have been accumulating at a dramatic rate. Inter-disciplinary research provides a new generation of computation tools in knowledge discovery and data management is in great demand. In this study, an expert-guided decision tree construction strategy is proposed to offer an user-oriented knowledge discovery environment. The strategy allows experts, based on their expertise and/or preference, to override inductive decision tree construction process. Moreover, by reviewing decision paths, experts could focus on subsets of data that may be clues to new findings, or simply contaminated cases.*

INTRODUCTION

Experts tend to execute the reasoning process with a series of rules. The rules are abstracted from basic principles and cases they have experienced in their fields. Two convenient ways to illustrate the rules could be a combination of if-then-else sentences, and a decision tree. Generally, these two forms of reasoning knowledge can be transformed back and forth. With the simple idea that decision making criteria can be derived from cases, scientists have developed different methods to inductively learn decision trees/rules from exemplars collected from different problem domains[1,2]. An emerging field: knowledge discovery in databases (KDD)[3], extends the scope of knowledge engineering research to extracting knowledge from data records collected for routine use. The stepwise process in KDD includes: defining goal(s); data collecting, cleaning, and reduction; data analyzing and hypothesis selecting; data mining; interpreting mined pattern(s); validating and acting upon discovered knowledge[4](Figure 1). Among them, data mining is the key step that focus on applying some specific algorithms for extracting patterns- in the form of a decision tree or set of decision rules- from data. The inductive learning paradigm is one of the best candidates for the data mining, because its end-product - decision trees/rules are the most comprehensive for human experts to review.

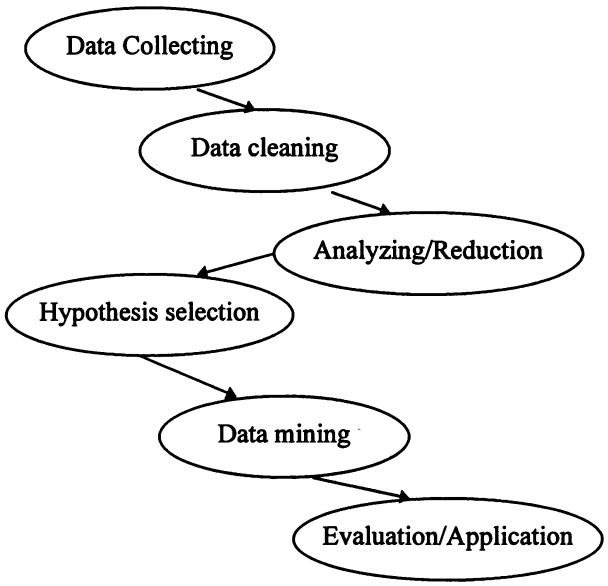

Figure 1. Stepwise process of Knowledge Discovery in Databases

When applying the KDD principles in biomedicine fields, several practical difficulties are evident:

- One may want to include cases that could cover all possible situations in the problem domain before starting the decision tree buildup. However, the decision about which cases should be in the training set is always debatable.

- There is no guarantee that the cases collected do not have noise. There is no any easy way to judge which case is "polluted", except by manual inspection by experts. When dealing with large amounts of data, manual inspection is not practical. Therefore, a decision tree generated from the contaminated data set would not truly reflect the domain knowledge.

1091—8280/97/$5.00

- The attributes selected to construct the decision trees could be inter-dependent. It is also possible that some less important attributes have not been included. The former condition would yield redundant computation load, whereas the latter one would still have many undetermined terminals in the finalized decision tree.

- A decision tree purely based upon inductive learning criteria would be biased by the training data characteristics. The decision path would sometimes violate "rules of thumb". In this case experts may be reluctant to carefully review the decision making structure. Therefore, the evaluation they did would be just statistical judgment of the performance of the learning data set and/or test data set.

In this article, an expert-guided decision tree construction strategy is proposed to help physicians organize reasoning processes from the evidence of collected cases. The basic principle is to combine experts' precedence in decision making with the computation power of computerized inductive learning algorithms. Since the derived decision trees/rules partially follow experts' reasoning, it is easier for experts to trace the knowledge patterns. While evaluating the decision paths, experts can also review the cases to which the paths are associated. When encountering an unfamiliar path, the experts can judge whether this is a clue to new knowledge, or is caused by invalid data. The data cleaning process can be used implicitly in this fashion. The iterative processes: decision tree build-up, decision path criticizing, data cleaning, predicting attributes revision, would be executed until a satisfactory decision tree is built. This finalized decision tree also serves as an indexing structure for learning cases partitioning. Similar case retrieval can be easily achieved by finding what terminal that new case has fallen into.

FUNDAMENTALS OF INDUCTIVE LEARNING

Early experiments implementing concept learning systems, conducted by Hunt, et. al.[5], provided a generic concept about how a decision tree is constructed. The divide and conquer tree constructing process suggested, but did not provide an optimal approach in finding a compact decision tree. Exhaustive exploration of all possible tree structures from given exemplars is necessary in order to find the simplest construct, yet still predictive decision tree. Until greedy algorithms were later created[6,7], searching for the simplest tree is a major time consuming task. One of these greedy tree construction algorithms, proposed by Quinlan et. al., uses the concept of entropy to represent information in data sets. The average amount of information needed to identify the classes in S is expressed as,

$$Info(S) = -\sum_{j=1}^{k} \frac{freq(Cj,S)}{|S|} \times \log_2\left(\frac{freq(Cj,S)}{|S|}\right) \quad \text{.....(1)}$$

where $\frac{freq(Cj,S)}{|S|}$ is the probability of class Cj in data set S.

Each predictor could be a candidate test to split the data set into subsets. After applying similar information measurement to n split subsets based on test X, the total information required to identify the classes is

$$Info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times Info(T_i) \quad \text{.....(2)}$$

Thus, the information gained by partitioning T with the test X can be measured as

$$gain(X) = Info(T) - Info_X(T) \quad \text{...(3)}$$

Using the same information measurement, one may get potential information generated by splitting T into n subset with test X. This split information is defined as,

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right) \quad \text{...(4)}$$

where $|T|$ and $|T_i|$ are case number in data set T and subset T_i, respectively.

By using the *split info(X)* as a normalization factor, the *gain ratio(X)* can be calculated as,

$$gain\ ratio(X) = gain(X)/split\ info(X) \quad \text{..(5)}$$

A test X that yields maximum gain ratio will be chose for data set partitioning. The same criteria will then apply to each split subset. The iterative divide and conquer process executes until no further split is

required. This gain ratio inductive learning criterion is known as C4.5[7]. Though the inductive learning algorithms offer an automatic decision trees/rules generation mechanism, the outcome is biased by the learning data provided. A revised approach dominated by experts is necessary to insure the generality of the decision making scheme.

EXPERT-GUIDED DECISION TREE

Biomedicine is sometimes referenced as a weak theory domain, in which a large part of the reasoning knowledge is vague and described differently by various experts. Though the entry points in reviewing a case among different experts could not be the same, the conclusion should be similar. The precedence factors related to outcome from different expert's viewpoint also varies. The proposed expert-guided decision tree construction strategy allows experts to enter an ordered list of predictors, which they believed to be significant in discriminating cases at beginning stage of decision tree. For those predictors not on the list, the partition procedure follows the inductive learning method.

After a derived decision tree is constructed, all the learning cases would be partitioned along the decision paths in the tree. The experts can then review the reasoning process that form the decision tree. If any conclusion suggested from the tree is in conflict with their idea, the group of cases in that terminal can be further reviewed by experts. Therefore, based on their expertise, whether there is a new finding, or a error path caused by contaminated cases, could be easily screened out.

Another refinement in decision tree construction is the adjustment of predictor list. Databases designed for routine clinical use would allocate all possible parameters to describe patient profiles from several aspects. When attempting to explore a specific outcome model from this kind of databases, all the outcome related predicting attributes listed by experts would be incomplete at beginning stage. When there are too many different categories of outcome in a tree terminal, the possible reason is that some less significant predictors are not included in this decision model. The experts can than revise the predictor set and re-generate a decision tree.

Under a perfect condition, the selected cases and predictors can truly reflect all the possible conditions in the problem domain. The decision model derived would not yield any vague outcome. In biomedicine, this situation rarely happens. Most of time, even with a search path that reaches a terminal of the decision tree, the exact outcome is still not available. This is due to complex nature of this field, such that the predictors included in decision tree are not, and will not be, able to completely describe a case. To deal with this situation, instead of trying to find out an ultimate predictor set, we can use the learning cases that have been categorized in the tree terminal, as a basis for later case match processing. In this way, a case to be classified will first follow a decision path based on its values of attributes that formed the decision tree. The k-Nearest Neighbors matching will be used to find k the most similar cases to the new cases from the basis. The majority of these k similar cases will suggest a potential outcome. The user can certainly override this result by reviewing retrieved similar cases and then make a final conclusion. The expert-guided decision tree construction strategy offers several advantages:

- It helps experts to organize their reasoning process with the evidence provided by collected cases.
- By presenting the proposed decision paths, it allows experts to screen out unwanted contaminated cases, to refine the predictor set, and to explore new knowledge.
- The hierarchical tree structure works as an indexing scheme for learning cases partition. This reduces the number of cases which go to case matching process.

In another sense, for a new case classification, the decision tree is used to form a primary hypothesis. The hypothesis confirmation can be finalized by human expert, or by later case-based reasoning process.

IMPLEMENTATION

The expert-guided decision tree construction strategy has been implemented on a Pentium-based Windows NT workstation. Software development uses Microsoft Visual Basic in junction with Open Data Base Connectivity(ODBC) drivers. With the easily manipulated graphical user interface(GUI), users can browse the content of any database that is ODBC supported, select a target outcome, setup a list of predictors, arrange the precedence of key predictors, and then launch the decision tree construction

process. The derived decision tree is represented in a TreeView customer control. The embedded node expansion and collapse controls in the TreeView allow users to easily focus on part of decision paths. A log file is created to record all the actions a user order during the whole knowledge discovery process.

PRELIMINARY RESULT

To prove the concept of proposed strategy, a database that has 286 breast cancer cases, provided by M. Zwitter and M. Soklic , Institute of Oncology, University Medical Centre, Ljubljana, Yugoslavia, has been used to compare the relative performance between expert-guided and non-expert-guided decision trees. Past usage of this data set shows 65% to 72% accuracy with different classification methods[2]. One of two outcome classes, no-recurrence-event and recurrence-event, has been marked for each case. Each case is described by nine predictors. The predictors and possible values are listed in Table 1. Ten cross-validation bootstraps, each with 200 (70%) training cases and 86 (30%) testing cases, were used for the performance evaluation. Three major predictors, *degree-malignancy*, *tumor-size*, and *invasive-node-counts* have been pointed out by a pathologist. These three attributes were put on the preference list for the proposed decision tree construction strategy. Part of the decision tree is listed in Figure 2. Table 2 reports the performance of decision trees with different experimental setups.

DISCUSSION AND SUMMARY

The preliminary results indicate that the expert-guided approach's performance is comparable to the optimal inductive learning approach. Including the k-NN case matching algorithm improves the performance of induced decision making. The proposed strategy is an effective tool for extracting knowledge from databases in conjunction with expert experience. The inclusion of an expert in the design should increase acceptability by physicians. Further studies to validate this approach in clinical practice are listed below:

- Employ advanced Inductive Logic Programming (ILP) techniques in decision trees/rule induction.
- Develop intelligent tree pruning paradigms to simplify tree structure, and thus reduce the experts' load when they are reviewing decision paths.
- Exploit Case-Based Reasoning principles in the final decision making stage.
- Integrate and evaluate the proposed strategy with Vanderbilt University Perioperative Information Management System (VPIMS), in performing clinical knowledge discovery in risk assessment and adverse outcome prediction.

REFERENCES

1. Quinlan, J. R. Decision Trees and Decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics* 20: 339-346, 1990.
2. Dezelic, G. and J. Kern. Inductive Learning as a Method for Medical Decision Making. *Proceedings of Medical Informatics Europe 1991* 327-331, 1991.
3. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to knowledge Discovery in Databases. *AI Magazine* 37-54, 1996.
4. Brachman, R. and T. Anand. The Process of Knowledge Discovery in Databases: Ahuman-Centered Approach. In: *Advances in Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Menlo Park, Calif. AAAI Press, 1996, 37-58.
5. Hunt, E. B., J. Marin, and P. J. Stone. *Experiments in Induction.* New York: Academic Press, 1966.
6. Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classfication and Regression Trees.* Belmont, CA: Waddworth International Group, 1984.
7. Quinlan, J. R. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann, 1993.

Table 1. The possible values of each predictor from the breast cancer database

Attribute	Possible Values
Outcomes	no-recurrence-event, recurrence-event
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
node-caps	yes, no
deg-malig	1, 2, 3
breast	left, right
breast-quad	left_up, left_low, right_up, right_low, central
irradiat	yes, no

Table 2. Performance matrix of different experiment setups

	Decision Tree Construction strategy					
	with expert-guidance			without expert-guidance		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
with k-NN	71.4%	45.4%	82.0%	68.6%	41.5%	79.6%
without k-NN	58.5%	29.4%	70.5%	60.2%	36.5%	70.0%

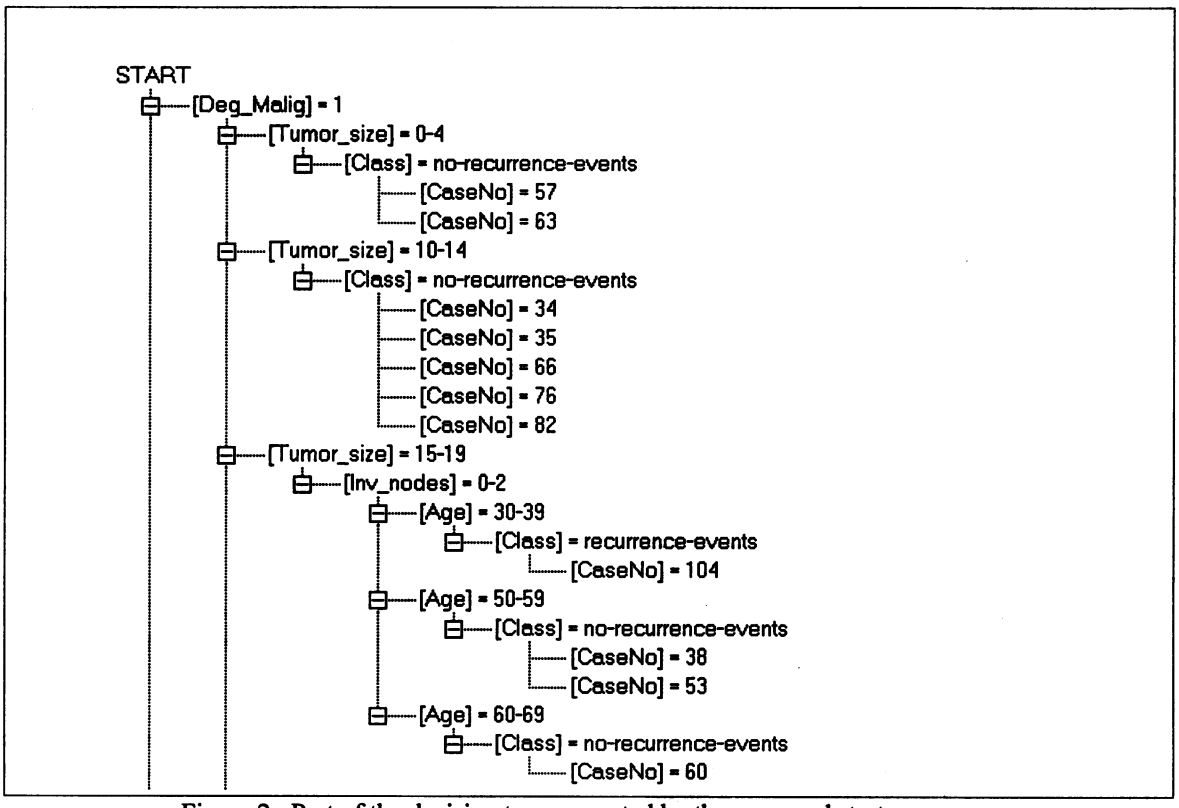

Figure 2. Part of the decision tree generated by the proposed strategy